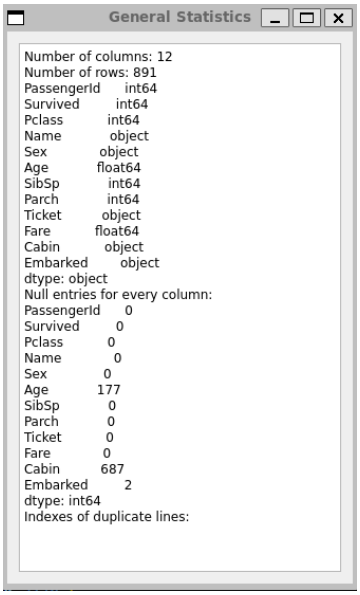
README

Partea 2 – PCLP3 – Titanic

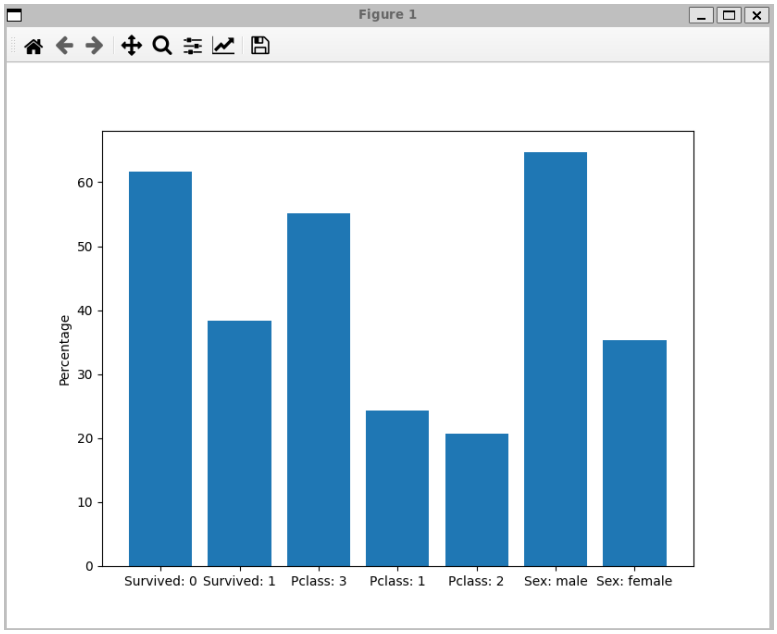
Cerinta 1

Datele cerute se obtin fie prin prelucrari ale fisierului train.csv, fie folosind data.info(). Am ales prima varianta pentru o mai mare manevrabilitate asupra rezultatului.



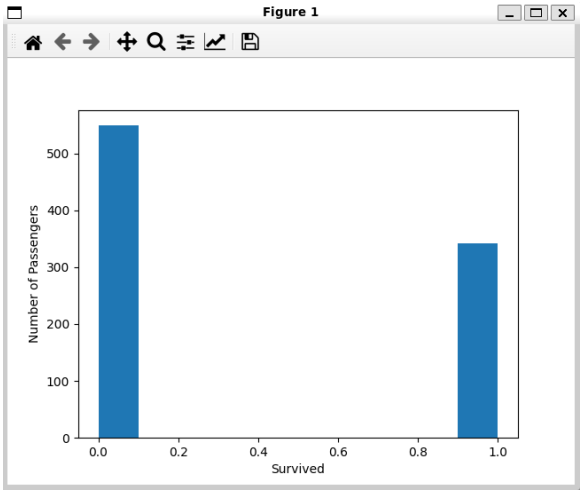
Cerinta 2

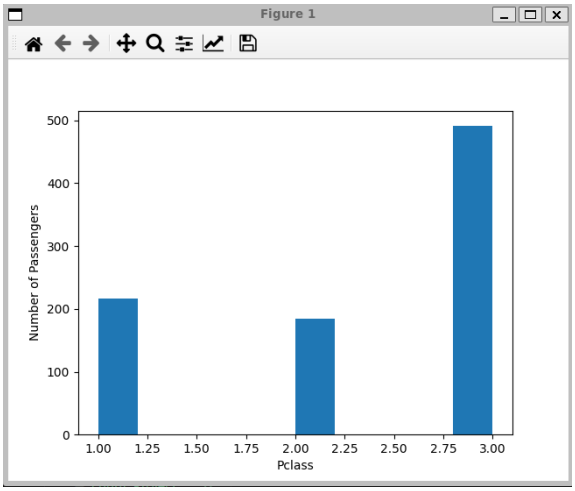
Procentele se obtin prin folosirea metodei value_counts(), impartirea rezultatului la numarul total de linii din data, apoi inmultind cu 100.

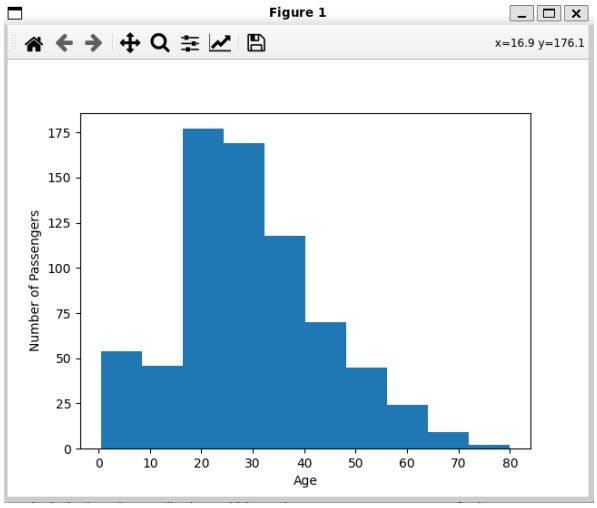


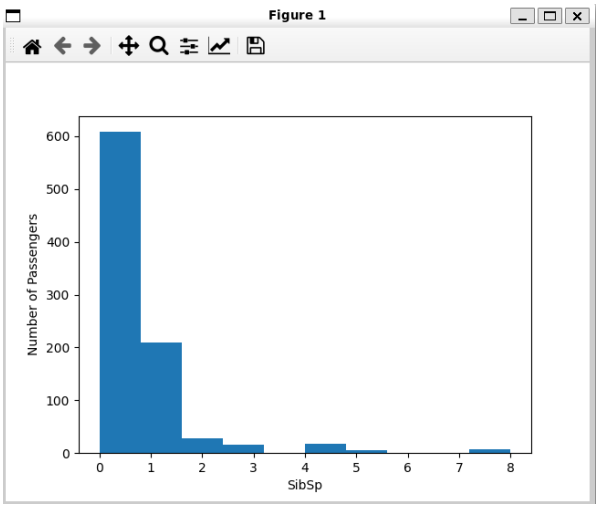
Cerinta 3

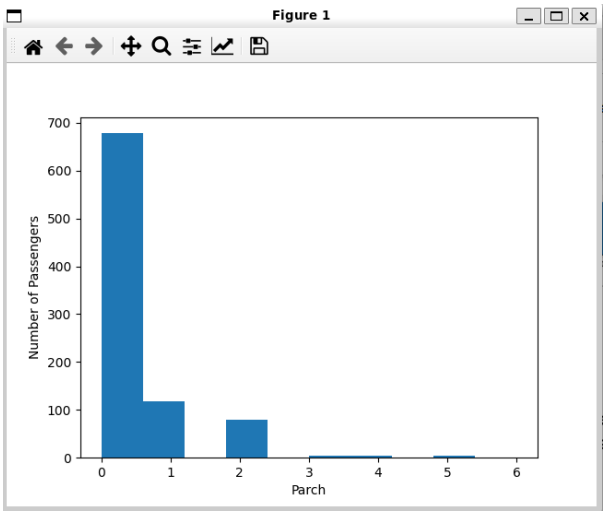
Histogramele se realizeaza prin folosirea functiei hist() din matplotlib, careia i se da ca argument coloana corespunzatoare din dataframe.

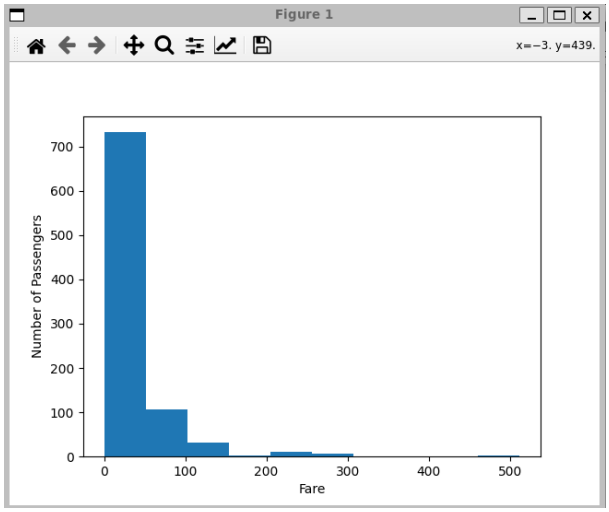






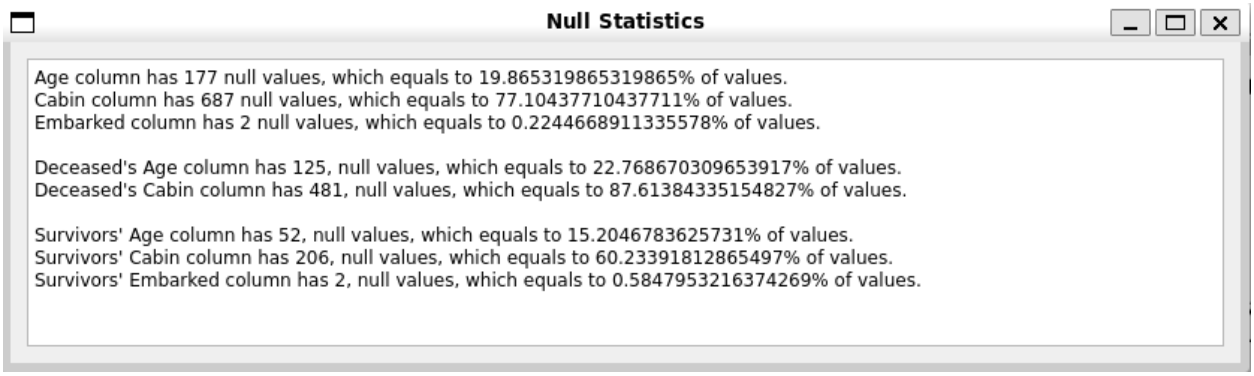






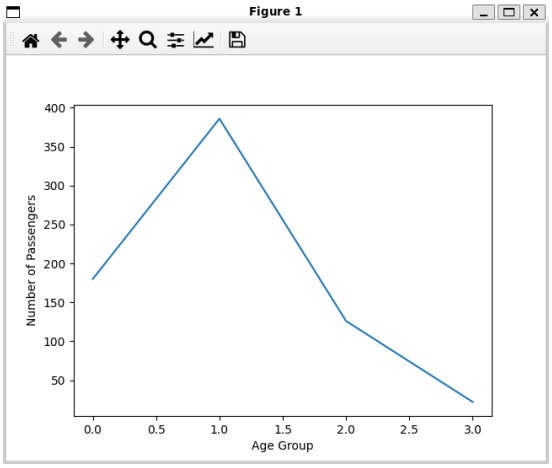
Cerinta 4

Numarul de elemente nule se afla folosind metoda isnull() urmata de metoda sum(). Procentele cerute se afla prin fragmentarea setului de date in functie de anumite conditii.



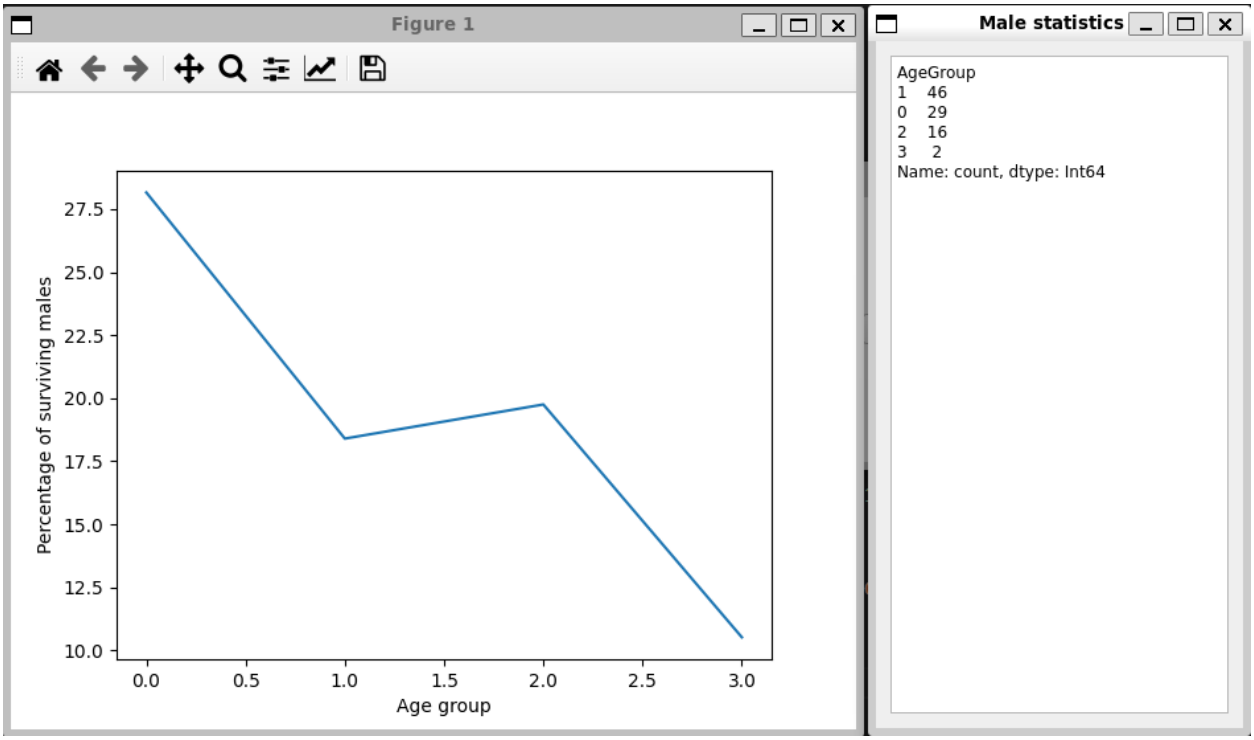
Cerinta 5

Coloana AgeGroup se obtine aplicand o functie lambda pe coloana Age, functie care scade 1 din varsta data (pentru ca marginile intervalelor sa fie incadrate corect), iar dupa impartirea intreaga la 20 regleaza rezultatul astfel incat acesta sa nu fie mai mic ca 0 (rezultat care se obtinea pentru bebelusii mai mici de 1 an) ori mai mare ca 3 (rezultat care se obtinea pentru o varsta initiala mai mare sau egala cu 81 de ani).



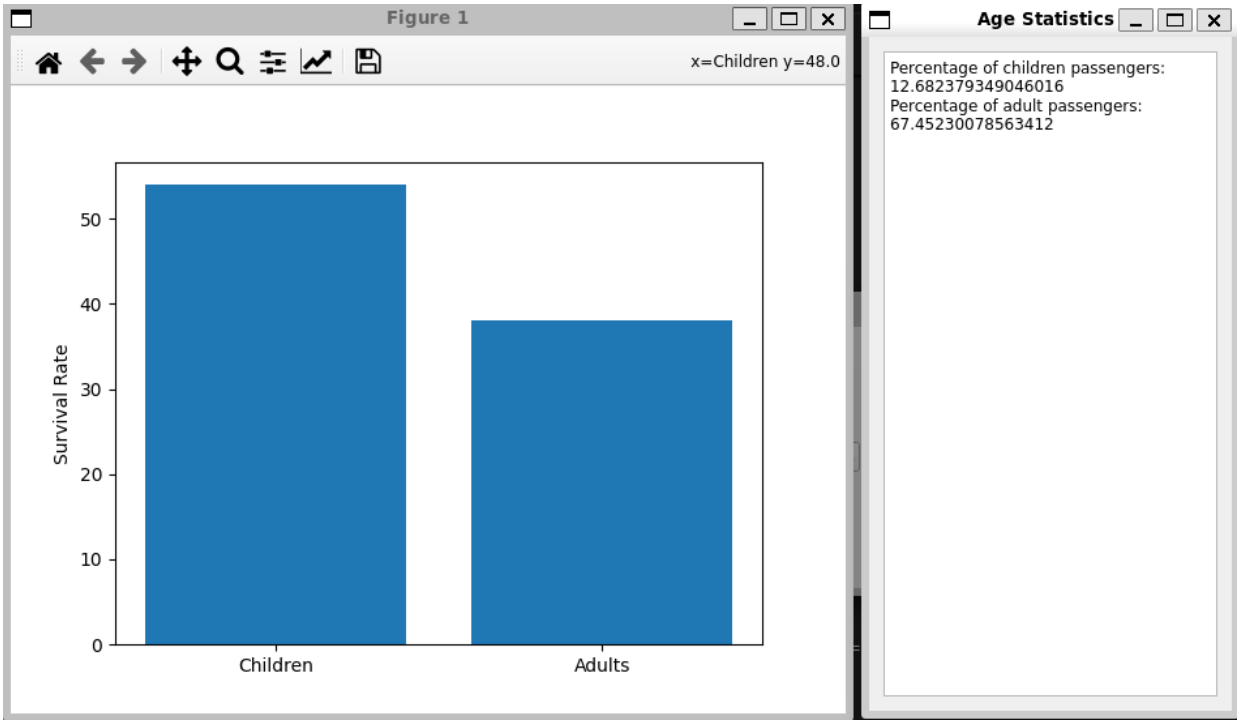
Cerinta 6

Functia nu prezinta noutati la capitolul prelucrare a setului de date. In cazul in care inainte de rulare nu s-a adaugat coloana AgeGroup, atunci aceasta va fi adaugata.



Cerinta 7

Se fac prelucrarile obisnuite pe cele doua sectiuni complementare in care impartim dataframe-ul (Age < 18 si Age >= 18). Motivul pentru care procentul de copii si cel de adulti nu insumeaza 100 este prezenta liniilor cu valori lipsa pentru coloana Age.

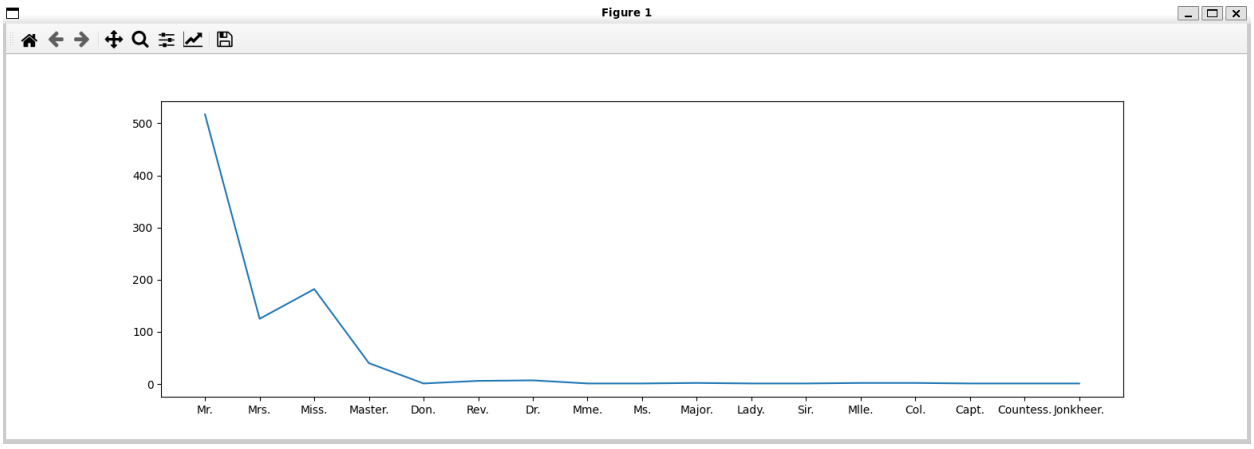


Cerinta 8

Pentru coloana Age valorile nule din dataframe vor fi inlocuite cu media de varsta pentru fiecare dintre cele doua clase (Survived si Deceased) folosind metoda loc(). In cazul celor doua coloane categoriale (Embarked si Cabin), valorile nule pentru fiecare clasa vor fi inlocuite cu cea mai frecventa valoare pentru clasa respectiva folosind metoda value_counts().

Cerinta 9

Se porneste de la doua array-uri continand titluri uzuale aplicabile unui singur sex. Cream un dictionar continand toate titlurile drept chei si avand drept valori numarul de persoane care detin titlul si au sexul corect. Acest lucru se realizeaza prin parcurgerea tuturor numelor din dataframe, iar pentru titlurile „unisex” se numara toate aparitiile titlului, fara a se mai verifica sexul.



Cerinta 10

Investigam influenta ipotetica a numarului de rude de pe vas asupra ratei de supravietuire folosind metoda corr(). Rezultatul scazut sugereaza ca legatura corelativa existenta intre cele doua statistici este una slaba, infirmand ca numarul de rude de pe vas a influentat considerabil rata de supravietuire. Din histograma putem trage concluzia pertinenta ca rata de supravietuire a scazut exponential pe masura ce numarul clasei (coloana Pclass) crestea, coborand de la peste 50% (in cazul Pclass = 1) pana la sub 25% (in cazul Pclass = 3).